Large-scale heterogeneous service systems with general packing constraints

Alexander L. Stolyar
Lehigh University
200 W. Packer Ave., Room 484
Bethlehem, PA 18015
stolyar@lehigh.edu

May 20, 2016

Abstract

A service system with multiple types of customers, arriving according to Poisson processes, is considered. The system is heterogeneous in that the servers also can be of multiple types. Each customer has an independent exponentially distributed service time, with the mean determined by its type. Multiple customers (possibly of different types) can be placed for service into one server, subject to "packing" constraints, which depend on the server type. Service times of different customers are independent, even if served simultaneously by the same server. The large-scale asymptotic regime is considered such that the customer arrival rates grow to infinity.

We consider two variants of the model. For the *infinite-server* model, we prove asymptotic optimality of the *Greedy Random* (GRAND) algorithm in the sense of minimizing the weighted (by type) number of occupied servers in steady-state. (This version of GRAND generalizes that introduced in [15] for the homogeneous systems, with all servers of same type.) We then introduce a natural extension of GRAND algorithm for *finite-server* systems with blocking. Assuming subcritical system load, we prove existence, uniqueness, and local stability of the large-scale system equilibrium point such that no blocking occurs. This result strongly suggests a conjecture that the steady-state blocking probability under the algorithm vanishes in the large-scale limit.

Keywords: Queueing networks, Stochastic bin packing, Heterogeneous service systems, Packing constraints, Blocking, Loss, Greedy random (GRAND) algorithm, Fluid limit, Cloud computing

AMS Subject Classification: 90B15, 60K25

1 Introduction

We consider a heterogeneous service system where servers can be of multiple types. There are also multiple types of customers, each arriving according to an independent Poisson process. Each customer has an independent exponentially distributed service time, with the mean determined by its type. Multiple customers (possibly of different types) can be placed for service into one server, subject to "packing" constraints, which depend on the server type. Service times of different customers are independent, even if served simultaneously by the same server. Such a system arises, for example, as a model of dynamic real-time assignment of virtual machines ("customers") to physical host machines ("servers") in a network cloud [6], where typical objectives may be to minimize the number of occupied (non-idle) hosts or to minimize blocking/waiting of virtual machines. In this paper we consider two variants of the system, and study their properties in the large-scale asymptotic regime, when the customer arrival rates (and then the number of occupied servers) are large.

The first variant of the system is such that there is an infinite "supply" of servers of each type. Each arriving customer is assigned to a server immediately upon arrival. The asymptotic regime is considered such that the customer arrival rates grow in proportion to a scaling parameter $r \to \infty$. Each server type s is assigned a weight ("cost") γ_s, and the objective is to minimize the weighted number ("total cost") of occupied servers in steady-state. We prove that a generalized version of the *Greedy Random* (GRAND) algorithm, introduced in [15] for a homogeneous system (with one server type), is asymptotically optimal, in the sense described below in this paragraph. The basic idea of GRAND is to assign an arriving customer of a given type i to a server chosen randomly uniformly among servers available to it, i.e. those servers where a type i customer can be added without violating packing constraints. A particular GRAND algorithm that we consider for the infinite server system, which is labeled GRAND($\boldsymbol{a}Z$), is as follows. There is a parameter $a_s > 0$ for each server type s; $\boldsymbol{a} = (a_s)$ is the vector with components a_s. An arriving customer picks uniformly at random an available server among all currently occupied servers plus designated numbers $a_s Z$ of idle servers (called "zero-servers") of each type s, where Z is the current total number of all customers. (GRAND(aZ) algorithm of [15] is a special case of GRAND($\boldsymbol{a}Z$), with single parameter $a > 0$, because there is only one server type.) *GRAND($\boldsymbol{a}Z$) achieves optimality if we first take the limit of system stationary distributions as $r \to \infty$, and then take the limit on $a_s = \alpha^{\gamma_s} \downarrow 0$, with common parameter $\alpha \downarrow 0$.* (We believe that a stronger form of asymptotic optimality, when only the limit $r \to \infty$ is taken, holds for a different version of GRAND, with the number of zero-servers of type s equal to $Z^{(p-1)\gamma_s+1}$, where parameter $p < 1$ is close to 1. See Conjecture 4 at the end of Section 2.2.)

It is important to emphasize that GRAND($\boldsymbol{a}Z$) achieves asymptotic optimality *without utilizing any knowledge of the system structural parameters*. Namely, the algorithm need not "know" the server types or exact states of the currently occupied servers. All it needs to know about each currently occupied server is whether or not it can "accept" an additional customer of type i, for each i. Note that the setting of the algorithm parameters a_s, that achieves asymptotic optimality, depends only on the weights γ_s, which are the parameters of the objective (as opposed to system parameters). One of the key qualitative insights of [15] was the surprising fact that an algorithm as simple as GRAND can be asymptotically optimal. The fact that an appropriately generalized, but still extremely simple, version of GRAND is optimal for in a heterogeneous system, is still more surprising.

The second variant is a system with finite size pools of servers of each type. Each arriving customer can be either immediately assigned to a server or immediately blocked (in which case it leaves the system without receiving service). The asymptotic regime is such that both the arrival rates and the server pool sizes scale in proportion to parameter $r \to \infty$. We consider a different version of the GRAND algorithm, labeled GRAND-F, which simply assigns each arriving customer randomly uniformly to any available to it server in the system, and blocks the customer if there are no such available servers. We study the dynamics of the fluid paths (obtained by "fluid" scaling and then the $r \to \infty$ limit). Assuming the system is subcritically loaded, we prove existence, uniqueness and local stability of a system equilibrium point, such that there is no blocking. These results strongly suggest a conjecture that GRAND-F is asymptotically optimal in that, under subcritical load, the limit of the system stationary distributions is concentrated on the equilibrium point described above, and therefore *the steady-state blocking probability vanishes in the $r \to \infty$ limit*. We note that the equilibrium point local stability property is stronger than a typical "fixed point" argument, based on the assumption of asymptotic independence of server states (or, "independence ansatz," in the terminology of [2, 3]). The fixed point argument allows one to characterize (and then possibly derive) the limit of the stationary distributions, assuming the ansatz holds. If the ansatz is proved, this of course proves the limit of the stationary distributions. If the ansatz is *not* proved, the fixed point argument is equivalent to the property that the equilibrium point is an invariant point of the fluid paths. The local stability of the equilibrium point that we prove, is a stronger property than just its existence and invariance, and therefore it provides a stronger support for the asymptotic optimality conjecture. (The relation between the local stability and the fixed point argument is discussed in detail in Section 5.1.)

We want to emphasize that the packing constraints that we consider are extremely general. (They are of the same kind as those in [13–15]; we additionally allow them to depend on the server type.) In particular, they are far more general than *vector packing* constraints. Vector packing refers to the situation when a server has the corresponding resource-vector, giving the amounts of resources of different types that it possesses; for each customer type there is the requirement-vector, giving the resource requirements of one customer; the constraint is that the sum of the requirement-vectors of the customers placed into a server cannot exceed its resource-vector. Packing of virtual machines into physical machines in a network cloud [6] is an example of vector packing.

Finally, we note that GRAND-F can be very efficiently implemented via a "pull-based" mechanism (see [16] and references therein), which has a very low signaling message exchange rate between the "router" and the servers. In fact, GRAND-F algorithm can be viewed as an extension of PULL algorithm [16] to service systems with packing constraints. (This is discussed in more detail in Remark 6 in Section 2.3.)

1.1 Related previous work

As mentioned above, the main practical motivation for our model is the problem of real-time dynamic assignment of virtual machines (VM) to physical host machines (PM) in a network cloud. (A general discussion of the issues that arise in this application can be found in [6].) Since multiple VMs can simultaneously occupy (be "packed into") same PM, this naturally leads to bin packing type models. There is an extensive literature on the classical bin packing (see, e.g., [1, 4, 8] for reviews and recent results), where each "item" (customer) once placed into a "bin" (server) stays in that bin forever. However, the dynamic VM-to-PM assignment problem is such that each VM (customer) leaves its PM (server), and the system, after its service is completed. This in turn naturally leads the models that we consider, i.e. service systems with packing constraints at the servers.

The infinite-server variant of our model is a generalization of the homogeneous (one server type) model studied in [13–15], which focused on the problem of minimizing the number of occupied servers in steady-state. In particular, GRAND algorithm was proposed and shown to be asymptotically optimal in [15]. (Papers [13, 14] have studied a different algorithm, which needs to know the structure of packing constraints and to use the exact current states of all servers.) Our model allows, in addition, multiple server types and we consider a more general problem of minimizing the weighted number of servers; the analysis of this variant of our model is a generalization of that in [15]. A homogeneous infinite-server model, specialized to vector packing constraints, was also considered in [5], where a randomized version of Best Fit algorithm was proved asymptotically optimal.

The finite-server variant of our model is related to the model in recent paper [19], which considers blocking in a homogeneous system, specialized to one-dimensional (single resource) vector packing constraints. (In [19] all servers are of the same type, and the term *heterogeneous* refers to multiple customer types, which our model also allows. So, in our terminology, the system in [19] is homogeneous.) The algorithm in [19] is of the *power-of-d-choices* type [2, 3, 12, 18], namely each arriving customer goes to the server which has the largest amount of unused resource, out of the d servers chosen uniformly at random. The paper uses a fixed point argument (independence ansatz) to derive the form of the equilibrium point, which is conjectured to be the asymptotic limit of the system steady-state. (In addition, the paper derives some performance bounds.) Of course, the equilibrium point under the power-of-d-choices algorithm is different from that under our GRAND-F algorithm. It is such that the blocking probability does *not* (and cannot be expected to) vanish in the limit. Therefore, the relation between the power-of-d-choices algorithm and GRAND-F for the systems with packing constraints, is analogous to the relation between power-of-d-choices and PULL algorithm [16] for service systems without packing, where the blocking (or waiting) probability vanishes under PULL, but not under the power-of-d-choices. (GRAND-F can be viewed as an extension of PULL algorithm to systems with packing constraints. See Remark 6 in Section 2.3.)

Papers [10, 11] consider a homogeneous finite-server system with queues (and no blocking), and focus on the system stability (or, throughput maximization). In [7] a heterogeneous finite-server system is considered, with the objective of minimizing maximum load across server pools; the algorithms proposed in [7] essentially treat the system as an infinite-server one. The algorithms in [7, 10, 11] are completely different from the variants of GRAND algorithm studied in this paper.

1.2 Layout of the rest of the paper

Basic notation used throughout the paper is given in Section 1.3. The model and the main results are stated in Section 2. The basic structure of the system, common to both variants, is given in Section 2.1. The infinite-server system, GRAND($\boldsymbol{a}Z$) algorithm and the main results for it (Theorems 2 and 3) are presented in Section 2.2. Section 2.3 defines the finite-server system, GRAND-F algorithm, and states the main result for it informally in Proposition 8 (with formal statements given later in Lemmas 15 and 16). Sections 3 and 4 contain proofs of the infinite-server/GRAND($\boldsymbol{a}Z$) results, while Section 5 contain those for finite-server/GRAND-F. Concluding remarks are given in Section 6.

1.3 Basic notation

Sets of real and real non-negative numbers are denoted by $\mathbb{R}$ and $\mathbb{R}_+$, respectively. We use bold and plain letters for vectors and scalars, respectively. The standard Euclidean norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\|$. Convergence $\boldsymbol{x} \to \boldsymbol{u} \in \mathbb{R}^n$ means ordinary convergence in $\mathbb{R}^n$, while $\boldsymbol{x} \to U \subseteq \mathbb{R}^n$ means convergence to a set, namely, $\inf_{\boldsymbol{u} \in U} \|\boldsymbol{x} - \boldsymbol{u}\| \to 0$. The i-th coordinate unit vector in $\mathbb{R}^n$ is denoted by $\boldsymbol{e}_i$. Symbol $\implies$ denotes convergence in distribution of random variables taking values in space $\mathbb{R}^n$ equipped with the Borel σ-algebra. The abbreviation *w.p.1* means convergence *with probability 1*. We often write $x(\cdot)$ to mean the function (or random process) $\{x(t),\ t \geq 0\}$. Abbreviation *u.o.c.* means *uniform on compact sets* convergence of functions. The cardinality of a finite set $\mathcal{N}$ is $|\mathcal{N}|$. Indicator function $I\{A\}$ for a condition A is equal to 1 if A holds and 0 otherwise. $\lceil \xi \rceil$ denotes the smallest integer greater than or equal to ξ, and $\lfloor \xi \rfloor$ denotes the largest integer smaller than or equal to ξ. For a finite set of scalar functions $f_n(t),\ t \geq 0,\ n \in \mathcal{N}$, a point t is called *regular* if for any subset $\mathcal{N}' \subseteq \mathcal{N}$ the derivatives $\frac{d}{dt} \max_{n \in \mathcal{N}'} f_n(t)$ and $\frac{d}{dt} \min_{n \in \mathcal{N}'} f_n(t)$ exist.

2 Model and main results

In this section we formally define the two variants of the model with heterogeneous servers, and state our main results for them. The first variant is a generalization of the infinite-server model in [13–15] in that we allow different types of servers, as opposed to just one type. The number of servers of each type is infinite and there is no blocking of arriving customers. For this version of the model the underlying objective is to minimize the weighted number of occupied servers in steady-state. The second variant is the model with different server types, but with finite number of servers of each type. If an arriving customer cannot be immediately assigned to some server in the system, it is blocked. In such a system, the underlying objective is to minimize blocking. Before defining these two variants of the model, in the next subsection we define the basic structure of the system (most importantly the server packing constraints), which is common for both model variants.

2.1 Heterogeneous servers. Packing constraints

We consider a service system with I types of customers, indexed by $i \in \{1, 2, \ldots, I\} \equiv \mathcal{I}$. The service time of a type-i customer is an exponentially distributed random variable with mean $1/\mu_i$. All customers' service times are mutually independent. There are S types of servers, indexed $s \in \{1, 2, \ldots, S\} \equiv \mathcal{S}$, and infinite "supply" of servers of each type. A server of each type can potentially serve more than one customer simultaneously, subject to the following very general packing constraints. We say that a vector $\boldsymbol{k} = (k_1, \ldots, k_I; s)$ with non-negative integer $k_i,\ i \in \mathcal{I}$, and $s \in \mathcal{S}$ is a server *configuration*, if a type s server can simultaneously serve a combination of customers of different types given by the values k_i. A configuration $\boldsymbol{k}$ with specific value of s is a type s server configuration. For any s, there is a finite set of all allowed type s server configurations, denoted by $\bar{\mathcal{K}}^s$. We assume that $\bar{\mathcal{K}}^s$ satisfies a natural *monotonicity* condition: if $\boldsymbol{k} \in \bar{\mathcal{K}}^s$, then all "smaller" configurations $\boldsymbol{k}' = (k'_1, \dots, k'_I; s)$, i.e. such that $k'_i \le k_i$ for all i, belong to $\bar{\mathcal{K}}^s$ as well. Without loss of generality, assume that for each i, $(\boldsymbol{e}_i; s) \in \bar{\mathcal{K}}^s$ for at least one s, where $\boldsymbol{e}_i$ is the i-th coordinate unit vector (otherwise, type-i customers cannot be served at all). By convention, for any s, vector $\boldsymbol{0}^s \equiv (\boldsymbol{0}; s) \in \bar{\mathcal{K}}^s$, where $\boldsymbol{k} = \boldsymbol{0}$ is the I-dimensional component-wise zero vector – this is the configuration of an empty type s server. We denote by $\mathcal{K}^s = \bar{\mathcal{K}}^s \setminus \{\boldsymbol{0}^s\}$ the set of type s server configurations *not* including the empty (or, zero) configuration. Denote by $\bar{\mathcal{K}} = \cup_s \bar{\mathcal{K}}^s$ and $\mathcal{K} = \cup_s \mathcal{K}^s$ the sets of all configurations and all non-zero configurations, respectively. In what follows, we use the following slightly abusive notations: for $\boldsymbol{k} \in \bar{\mathcal{K}}$, $\boldsymbol{k} + \boldsymbol{e}_i$ means vector $\boldsymbol{k}$ with k_i replaced by $k_i + 1$, and similarly for $\boldsymbol{k} - \boldsymbol{e}_i$.

An important feature of the model is that simultaneous service does *not* affect the service time distributions of individual customers. In other words, the service time of a customer is unaffected by whether or not there are other customers served simultaneously by the same server. A customer can be "added" to an empty or occupied server, as long as the packing constraints are not violated. Namely, a type i customer can be added to a server of type s whose current configuration $\boldsymbol{k} \in \bar{\mathcal{K}}^s$ is such that $\boldsymbol{k} + \boldsymbol{e}_i \in \mathcal{K}^s$. When the service of a type-i customer by a server in configuration $\boldsymbol{k}$ is completed, the customer leaves the system and the server's configuration changes to $\boldsymbol{k} - \boldsymbol{e}_i$.

2.2 Infinite-server system

In this section we define the infinite-server system, the proposed generalized GRAND(aZ) assignment (or packing) algorithm, and state the asymptotic optimality results for this algorithm.

We consider a system, as described in Section 2.1, in which there is an infinite "supply" of servers of each type $s \in \mathcal{S}$. Customers of type i arrive as an independent Poisson process of rate $\Lambda_i > 0$; these arrival processes are independent of each other and of the customer service times. Each arriving customer is immediately placed for service in one of the servers, as long as packing constraints are not violated.

Denote by $X_{\boldsymbol{k}}$ the number of servers in configuration $\boldsymbol{k} \in \mathcal{K}^s$. The system state is then the vector $\boldsymbol{X} = \{X_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K}\}$.

A *placement algorithm* (or packing rule) determines where an arriving customer is placed, as a function of the current system state $\boldsymbol{X}$. Under any well-defined placement algorithm, the process $\{\boldsymbol{X}(t), t \ge 0\}$ is a continuous-time Markov chain with a countable state space. It is easily seen to be irreducible and positive recurrent: the positive recurrence follows from the fact that the total number $Y_i(t)$ of type-i customers in the system is independent from the placement algorithm, and its stationary distribution is Poisson with mean Λ_i/μ_i; we denote by $Y_i(\infty)$ the random value of $Y_i(t)$ in steady-state – it is, therefore, a Poisson random variable with mean Λ_i/μ_i. Consequently, the process $\{\boldsymbol{X}(t),\ t \ge 0\}$ has a unique stationary distribution; let $\boldsymbol{X}(\infty) = \{X_{\boldsymbol{k}}(\infty), \boldsymbol{k} \in \mathcal{K}\}$ be the random system state $\boldsymbol{X}(t)$ in stationary regime.

We are interested in finding a placement algorithm that minimizes the total weighted number of occupied servers $\sum_{\boldsymbol{k} \in \mathcal{K}} X_{\boldsymbol{k}}(\infty)$ in the stationary regime.

Consider the following generalization of the Greedy-Random (GRAND) algorithm, introduced in [15]. More specifically, it is a generalization of the special form of the algorithm, called in [15] GRAND(aZ).

Definition 1 (Greedy-Random (GRAND($\boldsymbol{a}Z$)) algorithm for heterogeneous infinite-server systems)**.** *The algorithm is parameterized by a vector $\boldsymbol{a} = (a_s,\ s \in \mathcal{S})$ of real numbers $a_s > 0$. Let $Z(t) = \sum_i \sum_{\boldsymbol{k}} k_i X_{\boldsymbol{k}}(t)$ denote the total number of customers in the system at time t. At any given time t, there is a designated finite set of $X_{\boldsymbol{0}^s}(t) = \lceil a_s Z(t) \rceil \ge 0$ empty type s servers, called* s-zero-servers.
A new customer, say of type i, arriving at time t is placed into a server chosen randomly uniformly among those zero-servers (of any type s) and occupied servers, where it can still fit. In other words, the total number

of servers available to a type-i arrival at time t is

$$X_{(i)}(t) \doteq \sum_{\boldsymbol{k} \in \bar{\mathcal{K}}:\ \boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} X_{\boldsymbol{k}}(t) \equiv \sum_{s:\ \boldsymbol{e}_i \in \mathcal{K}^s} \left[X_{\boldsymbol{0}^s}(t) + \sum_{\boldsymbol{k} \in \mathcal{K}:\ \boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} X_{\boldsymbol{k}}(t) \right].$$

If $X_{(i)}(t) = 0$, the customer is placed into an empty server of any type s such that $\boldsymbol{e}_i \in \mathcal{K}^s$.

The GRAND($\boldsymbol{a}Z$) algorithm is easily implementable. (A detailed discussion of the implementation issues of the GRAND algorithm is given below in Remark 6, in the context of finite-server systems.)

We now define the asymptotic regime. Let $r \to \infty$ be a positive scaling parameter. More specifically, assume that $r \geq 1$, and r increases to infinity along a discrete sequence. Customer arrival rates scale linearly with r; namely, for each r, $\Lambda_i = \lambda_i r$, where λ_i are fixed positive parameters. Let $(\boldsymbol{X}^r(t),\ t \geq 0)$, be the process associated with a system with parameter r, and let $\boldsymbol{X}^r(\infty)$ be the (random) system state in the stationary regime. (Note that we do *not* include the zero-server numbers $X^r_{\boldsymbol{0}^s}(t)$ into $\boldsymbol{X}^r(t) = \{X^r_{\boldsymbol{k}}(t),\ \boldsymbol{k} \in \mathcal{K}\}$.) For each i, denote by $Y_i^r(t) \equiv \sum_{\boldsymbol{k} \in \mathcal{K}} k_i X^r_{\boldsymbol{k}}(t)$ the total number of customers of type i. Since arriving customers are placed for service immediately and their service times are independent of each other and of the rest of the system, $Y_i^r(\infty)$ is a Poisson random variable with mean $r\rho_i$, where $\rho_i \equiv \lambda_i/\mu_i$. Moreover, $Y_i^r(\infty)$ are independent across i. Since the total number of occupied servers is no greater than the total number of customers, $\sum_{\boldsymbol{k} \in \mathcal{K}} X^r_{\boldsymbol{k}}(t) \leq Z^r(t) \equiv \sum_i Y_i^r(t)$, we have a simple upper bound on the total number of occupied servers in steady state, $\sum_{\boldsymbol{k} \in \mathcal{K}} X^r_{\boldsymbol{k}}(\infty) \leq Z^r(\infty) \equiv \sum_i Y_i^r(\infty)$, where $Z^r(\infty)$ is a Poisson random variable with mean $r \sum_i \rho_i$. Without loss of generality, from now on we assume $\sum_i \rho_i = 1$. This is equivalent to rechoosing the parameter r to be $r \sum_i \rho_i$.

The *fluid-scaled* process is $\boldsymbol{x}^r(t) = \boldsymbol{X}^r(t)/r$, $t \in [0, \infty)$. We also define $\boldsymbol{x}^r(\infty) = \boldsymbol{X}^r(\infty)/r$. For any r, $\boldsymbol{x}^r(t)$ takes values in the non-negative orthant $\mathbb{R}_+^{|\mathcal{K}|}$. Similarly, $y_i^r(t) = Y_i^r(t)/r$, $z^r(t) = Z^r(t)/r$, $x^r_{\boldsymbol{0}^s}(t) = X^r_{\boldsymbol{0}^s}(t)/r$ and $x^r_{(i)}(t) = X^r_{(i)}(t)/r$, for $t \geq 0$ and $t = \infty$. Since $\sum_{\boldsymbol{k} \in \mathcal{K}} x^r_{\boldsymbol{k}}(\infty) \leq z^r(\infty) = Z^r(\infty)/r$, we see that the random variables $(\sum_{\boldsymbol{k} \in \mathcal{K}} x^r_{\boldsymbol{k}}(\infty))$ are uniformly integrable in r. This in particular implies that the sequence of distributions of $\boldsymbol{x}^r(\infty)$ is tight, and therefore there always exists a limit $\boldsymbol{x}(\infty)$ in distribution, so that $\boldsymbol{x}^r(\infty) \implies \boldsymbol{x}(\infty)$, along a subsequence of r.

The limit (random) vector $\boldsymbol{x}(\infty)$ satisfies the following conservation laws:

$$\sum_{\boldsymbol{k} \in \mathcal{K}} k_i x_{\boldsymbol{k}}(\infty) \equiv y_i(\infty) = \rho_i, \quad \forall i, \tag{1}$$

and, in particular,

$$z_i(\infty) \equiv \sum_i y_i(\infty) \equiv \sum_i \rho_i = 1. \tag{2}$$

Therefore, the values of $\boldsymbol{x}(\infty)$ are confined to the convex compact $(|\mathcal{K}| - I)$-dimensional polyhedron

$$\mathcal{X} \equiv \{\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|} \mid \sum_s \sum_{\boldsymbol{k} \in \mathcal{K}^s} k_i x_{\boldsymbol{k}} = \rho_i,\ \forall i \in \mathcal{I}\}.$$

We will slightly abuse notation by using symbol $\boldsymbol{x}$ for a generic element of $\mathcal{X}$; while $\boldsymbol{x}(\infty)$, and later $\boldsymbol{x}(t)$, refer to random elements taking values in $\mathcal{X}$.

Also note that under GRAND(aZ), for any server type s, $x^r_{\boldsymbol{0}^s}(\infty) \implies x_{\boldsymbol{0}^s}(\infty) = a_s z(\infty) = a_s$, as $r \to \infty$.

The asymptotic regime and the associated basic properties (1) and (2) hold *for any placement algorithm.* Indeed, (1) and (2) only depend on the already mentioned fact that all $Y_i^r(\infty)$ are mutually independent Poisson random variables with means $\rho_i r$.

Let the server weights $\gamma_s > 0$, $s \in \mathcal{S}$, be fixed. (One can think of γ_s as the "cost" rate of using one type s server.) Consider the following problem of minimizing the weighted number of occupied servers, on the fluid scale: $\min_{\boldsymbol{x}\in\mathcal{X}} \sum_{s\in\mathcal{S}} \sum_{\boldsymbol{k}\in\mathcal{K}^s} \gamma_s x_{\boldsymbol{k}}$. It is a linear program:

$$\min_{\boldsymbol{x}\in\mathbb{R}_+^{|\mathcal{K}|}} \sum_{s\in\mathcal{S}} \sum_{\boldsymbol{k}\in\mathcal{K}^s} \gamma_s x_{\boldsymbol{k}}, \tag{3}$$

subject to

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, \quad \forall i. \tag{4}$$

Without loss of generality, assume that the weights are scaled so that $\gamma_1 = 1$. Denote by $\mathcal{X}^* \subseteq \mathcal{X}$ the set of optimal solutions of (3)-(4).

For future reference, we record the following observations and notation. Using the monotonicity of $\bar{\mathcal{K}}$, it is easy to check that if in the LP (3)-(4) we replace equality constraints (4) with the inequality constraints

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} \geq \rho_i, \quad \forall i, \tag{5}$$

the new LP (3),(5) has same optimal value, and its set of the optimal solutions $\mathcal{X}^{**}$ contains $\mathcal{X}^*$, or more precisely, $\mathcal{X}^* = \mathcal{X}^{**} \cap \mathcal{X}$. From here, using Kuhn-Tucker theorem, $\boldsymbol{x} \in \mathcal{X}^*$ if and only if there exists a vector $\boldsymbol{\eta} = \{\eta_i,\ i \in \mathcal{I}\}$ of Lagrange multipliers, corresponding to the inequality constraints (5), such that the following conditions hold:

$$\boldsymbol{x} \in \mathcal{X}, \tag{6}$$

$$\eta_i \geq 0, \quad \forall i \in \mathcal{I}, \tag{7}$$

$$\sum_i k_i \eta_i \leq \gamma_s, \quad \boldsymbol{k} \in \mathcal{K}^s, \tag{8}$$

$$\text{for } \boldsymbol{k} \in \mathcal{K}^s, \quad \text{condition } \sum_i k_i \eta_i < \gamma_s \text{ implies } x_{\boldsymbol{k}} = 0. \tag{9}$$

Vectors $\boldsymbol{\eta}$ satisfying (6)-(9) for some $\boldsymbol{x} \in \mathcal{X}$ are optimal solutions to the problem dual to LP (3),(5). They form a convex set, which we denote by $\mathcal{H}^*$; it is easy to check that $\mathcal{H}^*$ is compact.

For each parameter-vector $\boldsymbol{a}$ (as in the definition of GRAND($\boldsymbol{a}Z$) algorithm), denote

$$L^{(\boldsymbol{a})}(\boldsymbol{x}) = \sum_s \sum_{\boldsymbol{k}\in\mathcal{K}^s} x_{\boldsymbol{k}} \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/(e a_s)], \tag{10}$$

where $c_{\boldsymbol{k}} \doteq \prod_i k_i!$, $0! = 1$. Then for $\boldsymbol{k} \in \mathcal{K}^s$ we have

$$(\partial/\partial x_{\boldsymbol{k}}) L^{(\boldsymbol{a})}(\boldsymbol{x}) = \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/a_s]. \tag{11}$$

Note that if we adopt a convention that

$$(\partial/\partial x_{\boldsymbol{0}^s}) L^{(\boldsymbol{a})}(\boldsymbol{x})|_{x_{\boldsymbol{0}^s}=a_s} = 0, \tag{12}$$

then (11) is valid for $\boldsymbol{k} = \boldsymbol{0}^s$ and $x_{\boldsymbol{0}^s} = a_s$, which will be useful later.

The function $L^{(\boldsymbol{a})}(\boldsymbol{x})$ is strictly convex in $\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|}$. Consider the problem $\min_{\boldsymbol{x}\in\mathcal{X}} L^{(\boldsymbol{a})}(\boldsymbol{x})$. It is the following convex optimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}_+^{|\mathcal{K}|}} L^{(\boldsymbol{a})}(\boldsymbol{x}), \tag{13}$$

subject to

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, \quad \forall i. \tag{14}$$

Denote by $\boldsymbol{x}^{*,\boldsymbol{a}} \in \mathcal{X}$ its unique optimal solution. Using (11) it is easy to check that $x^{*,\boldsymbol{a}}_{\boldsymbol{k}} > 0$ for all $\boldsymbol{k} \in \mathcal{K}$. There exists a vector $\boldsymbol{\nu}^{*,\boldsymbol{a}} = \{\nu^{*,\boldsymbol{a}}_i,\ i \in \mathcal{I}\}$ of Lagrange multipliers for the constraints (14), such that $\boldsymbol{x}^{*,\boldsymbol{a}}$ solves problem

$$\min_{\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|}} L^{(\boldsymbol{a})}(\boldsymbol{x}) + \sum_i \nu^{*,\boldsymbol{a}}_i (\rho_i - \sum_{\boldsymbol{k} \in \mathcal{K}} k_i x_{\boldsymbol{k}}).$$

We see that $\log[x^{*,\boldsymbol{a}}_{\boldsymbol{k}} c_{\boldsymbol{k}}/a_s] - \sum_i \nu^{*,\boldsymbol{a}}_i k_i = 0$, $\boldsymbol{k} \in \mathcal{K}$. Therefore, $\boldsymbol{x}^{*,\boldsymbol{a}}$ has the product form

$$x^{*,\boldsymbol{a}}_{\boldsymbol{k}} = \frac{a_s}{c_{\boldsymbol{k}}} \exp\left[\sum_i k_i \nu^{*,\boldsymbol{a}}_i\right], \quad \boldsymbol{k} \in \mathcal{K}^s. \tag{15}$$

This in particular implies that the Lagrange multipliers $\nu^{*,\boldsymbol{a}}_i$ are unique and are equal to $\nu^{*,\boldsymbol{a}}_i = \log(x^{*,\boldsymbol{a}}_{\boldsymbol{e}_i}/a_s)$, by considering (15) for $\boldsymbol{e}_i$, $i \in \mathcal{I}$; note also that they can have any sign (not necessarily non-negative). Therefore, we obtain the following fact. *A point $\boldsymbol{x} \in \mathcal{X}$ is the optimal solution to (13)-(14) (that is $\boldsymbol{x} = \boldsymbol{x}^{*,\boldsymbol{a}}$) if and only if it has a product form representation (15) for some vector $\boldsymbol{\nu}^{*,\boldsymbol{a}}$.* (The 'only if' part we just proved, and the 'if' follows from Kuhn-Tucker theorem.)

Our main results on the asymptotic optimality of GRAND(aZ) algorithm for the system with infinite number of servers are the following Theorems 2 and 3.

Theorem 2. *Let the parameter vector $\boldsymbol{a}$ be fixed. Consider a sequence of systems under the GRAND($\boldsymbol{a}Z$) algorithm, indexed by r, and let $\boldsymbol{x}^r(\infty)$ denote the random state of the fluid-scaled process in the stationary regime. Then, as $r \to \infty$,*

$$\boldsymbol{x}^r(\infty) \implies \boldsymbol{x}^{*,\boldsymbol{a}}.$$

Theorem 3. *Suppose the parameter vector $\boldsymbol{a}$ itself depends on a single parameter $\alpha > 0$ as follows: $a_s = \alpha^{\gamma_s}, s \in \mathcal{S}$. Then, as $\alpha \downarrow 0$, $\boldsymbol{x}^{*,\boldsymbol{a}} \to \mathcal{X}^*$ and $(-\log\alpha)^{-1}\boldsymbol{\nu}^{*,\boldsymbol{a}} \to \mathcal{H}^*$.*

Theorems 2 and 3 show that GRAND($\boldsymbol{a}Z$) is asymptotically optimal in the sense that $\boldsymbol{x}^r(\infty)$ converges to the optimal set $\mathcal{X}^*$, if we first take the limit $r \to \infty$, and then take the limit $\alpha \downarrow 0$ with $a_s = \alpha^{\gamma_s}$.

It was proved in a recent paper [17] (which is posterior to this paper) that a stronger form of asymptotic optimality, when only the limit $r \to \infty$ is taken, is achieved by the following version of GRAND, called GRAND(Z^p). This is a GRAND algorithm with the number of zero-servers depending on Z as Z^p, where $p < 1$ is a parameter, which is sufficiently close to 1, but depends only on the packing constraints. GRAND(Z^p) can be informally interpreted as GRAND(aZ), with a being variable $a = Z^{p-1}$ rather than constant. This suggests that for the heterogeneous infinite-server system that we consider, the stronger form of asymptotic optimality should hold, if we make a_s variable, equal to $Z^{(p-1)\gamma_s}$. Specifically, we believe that the methods of [17] can be extended to prove the following fact.

Conjecture 4. *Consider the GRAND algorithm with the number of zero-servers of type s equal to $Z^{(p-1)\gamma_s+1}$, where parameter $p < 1$ is sufficiently close to 1, but depends only on the packing constraints (i.e., sets $\mathcal{K}^s$). Then, as $r \to \infty$, $d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \Rightarrow 0$, where $d(\boldsymbol{x}, U)$ is the distance from point $\boldsymbol{x}$ to set U.*

2.3 Finite-server system

We now consider a version of the system, where the number of servers of each type is finite. Namely, there is a finite number $H_s > 0$ of servers of type s. Customers of type i arrive as an independent Poisson process of rate $\Lambda_i > 0$ (and these processes are independent from the customer service times). Each arriving type i customer can be either immediately placed for service into one of the servers (subject to packing constraints) or immediately blocked, in which case it immediately leaves the system. If there is no server where an arriving customer can be placed, the customer is necessarily blocked.

Let $X_{\boldsymbol{k}}$ denote the number of servers in configuration $\boldsymbol{k} \in \mathcal{K}^s$ and the system state is the vector $\boldsymbol{X} = \{X_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K}\}$. (Same notation as for the infinite-server system.) Note that we do not include the numbers

$X_{\mathbf{0}^s}$ of empty servers of each type (i.e., s-zero-servers) into the state $\boldsymbol{X}$. However, those number are, of course, uniquely determined by $\boldsymbol{X}$, because at all times we have the conservation law

$$X_{\mathbf{0}^s} + \sum_{\boldsymbol{k} \in \mathcal{K}^s} X_{\boldsymbol{k}} = \sum_{\boldsymbol{k} \in \bar{\mathcal{K}}^s} X_{\boldsymbol{k}} = H_s, \ s \in \mathcal{S}.$$

In such a system, a placement algorithm (or packing rule) determines, depending on the current system state $\boldsymbol{X}$, whether or not an arriving customer is accepted (i.e., not blocked), and if so, into which server it is placed. (If there are no servers, where a customer can be placed, it is necessarily blocked.) Under any well-defined placement algorithm, the process $\{\boldsymbol{X}(t), t \geq 0\}$ is a continuous-time Markov chain with finite state space; it is easily seen to be irreducible and, therefore, ergodic, with unique stationary distribution. Let $\boldsymbol{X}(\infty) = \{X_{\boldsymbol{k}}(\infty), \boldsymbol{k} \in \mathcal{K}\}$ be the random system state $\boldsymbol{X}(t)$ in stationary regime. It is also easy to see that $Y_i(\infty)$ – the steady-state random number of all type i customers in the system – is stochastically dominated by that in the infinite-server system, i.e. by a Poisson random variable with mean Λ_i/μ_i.

For this system, the underlying objective is to minimize blocking in steady-state. We consider the following version of the Greedy-Random (GRAND) algorithm, for the finite-server systems. It will be labeled GRAND-F.

Definition 5 (GRAND-F)**.** *A new customer, say of type i, arriving at time t is placed into a server chosen randomly uniformly among all servers in the system where it can still fit. (The total number of servers available for a type i customer addition at time t is*

$$X_{(i)}(t) \doteq \sum_{\boldsymbol{k} \in \bar{\mathcal{K}}: \ \boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} X_{\boldsymbol{k}}(t). \ )$$

If there are no such available servers (i.e., $X_{(i)}(t) = 0$), the customer is blocked.

Remark 6. An implementation of GRAND-F algorithm only requires that the "router" (an entity, making an assignment decision for each arriving customer) knows which servers are currently available for an addition of a type i customer, for each $i \in \mathcal{I}$. The router does *not* need to know the exact configurations of the servers. Moreover, it does *not* even need to know the server types! Therefore, the router needs to maintain only I bits of information for each server. This in turn is easily achievable, for example, by using a *pull-based* mechanism, analogous to that used by the PULL algorithm proposed in [16] (in a different context, for systems without non-trivial packing constraints). A specific pull-based mechanism to work in conjunction with GRAND-F can be as follows.
(a) Upon a customer, say of type i, arrival, the router follows GRAND-F rule for choosing a server. If there are no available servers for type i, the customer is blocked and no further action is taken. If the customer is assigned to a server, the server availability state (I bits) is changed to indicate the unavailability to *any* customer type i.
(b) Each server, when its configuration changes, i.e. upon any customer arrival (assignment) or departure (service completion), sends a "pull-message" (I bits), containing its new availability state, to the router.
(c) When router receives a pull-message from a server, it updates its availability status accordingly. (In reality, to prevent router from using "obsolete" pull-messages, after assigning a customer to a server, router can use some short time-out for the server, during which the server is considered unavailable regardless of its availability state. Thus, when the time-out expires, the availability state of the server is that from the *latest* pull-message received from it. If the time-out is longer than the "round-trip" router-server-router message delay, then the latest pull-message from the server is generated upon the last customer assignment to it, or maybe later, upon departures that occurred after that.)
This mechanism is such that the rate of pull-messages in the system is very small, namely two pull-messages per each arriving customer. The low rate of communication between the router and the servers is a very important feature of pull-based algorithms, because in modern cloud based systems, the number of servers can be very large.
We also note that a key part of the PULL algorithm is the random uniform assignment of customers to available servers. Therefore, GRAND-F algorithm can be viewed as an extension of PULL algorithm to service systems with packing constraints.

We consider the asymptotic regime, where the arrival rates are increased linearly with a scaling parameter $r \to \infty$: $\Lambda_i = \lambda_i r$, where $\lambda_i > 0$ are fixed parameters. In addition, so do the server pool sizes H_s, namely, $H_s = h_s r$, where $h_s > 0,\ s \in \mathcal{S}$, are fixed parameters.

Let $\boldsymbol{X}^r(\cdot)$ be the process associated with a system with parameter r, and let $\boldsymbol{X}^r(\infty)$ be the (random) system state in the stationary regime. For each i, denote by $Y_i^r(t) \equiv \sum_{\boldsymbol{k}\in\mathcal{K}} k_i X_{\boldsymbol{k}}^r(t)$ the total number of customers of type i. As mentioned above, $Y_i^r(\infty)$ is stochastically dominated by a Poisson random variable with mean $r\rho_i$, where $\rho_i \equiv \lambda_i/\mu_i$. As before, without loss of generality, we assume $\sum_i \rho_i = 1$.

The *fluid-scaled* process is $\boldsymbol{x}^r(t) = \boldsymbol{X}^r(t)/r,\ t \in [0,\infty)$. We define $\boldsymbol{x}^r(\infty) = \boldsymbol{X}^r(\infty)/r$. Similarly, $y_i^r(t) = Y_i^r(t)/r$, $x_{\boldsymbol{0}^s}^r(t) = X_{\boldsymbol{0}^s}^r(t)/r$ and $x_{(i)}^r(t) = X_{(i)}^r(t)/r$, for $t \geq 0$ and $t = \infty$.

For any r, $\boldsymbol{x}^r(t)$ takes values in the compact set

$$\mathcal{X}^{\square} \equiv \{\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|} \mid \sum_{\boldsymbol{k}\in\mathcal{K}^s} x_{\boldsymbol{k}} \leq h_s,\ \forall s \in \mathcal{S}\}.$$

For any $\boldsymbol{x} \in \mathcal{X}^{\square}$, we denote $x_{\boldsymbol{0}^s} \equiv h_s - \sum_{\boldsymbol{k}\in\mathcal{K}^s} x_{\boldsymbol{k}},\ s \in \mathcal{S}$, and will sometimes use notation $\bar{\boldsymbol{x}} \equiv \{x_{\boldsymbol{k}},\ \boldsymbol{k} \in \bar{\mathcal{K}}\}$.

The sequence of distributions of $\boldsymbol{x}^r(\infty)$ is obviously tight, and therefore there always exists a limit $\boldsymbol{x}(\infty)$ in distribution, so that $\boldsymbol{x}^r(\infty) \implies \boldsymbol{x}(\infty)$, along a subsequence of r. The limit (random) vector $\boldsymbol{x}(\infty)$ satisfies the following property w.p.1.:

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}(\infty) \equiv y_i(\infty) \leq \rho_i, \quad \forall i. \tag{16}$$

The asymptotic regime and property (16) obviously hold for any placement algorithm, not just GRAND-F.

Consider the following subset of $\mathcal{X}^{\square}$:

$$\mathcal{X}^{\diamond} \equiv \{\mathcal{X} \in \mathcal{X}^{\square} \mid \sum_s \sum_{\boldsymbol{k}\in\mathcal{K}^s} k_i x_{\boldsymbol{k}} = \rho_i,\ \forall i \in \mathcal{I}\} \equiv \mathcal{X}^{\square} \cap \mathcal{X}.$$

We make the following

Assumption 7. *The system parameters λ_i, μ_i, $i \in \mathcal{I}$, and h_s, $s \in \mathcal{S}$, are such that the set $\mathcal{X}^{\diamond}$ in non-empty. Moreover, there exists $\boldsymbol{x} \in \mathcal{X}^{\diamond}$ such that $x_{\boldsymbol{0}^s} > 0$ for all s.*

This assumption means that, when the scaling parameter r is large, and we have $\rho_i r$ customers of each type i, it is possible to "pack" all of them into the system servers ($h_s r$ for each type s), so that a non-zero fraction of servers in each pool s remains idle. Recall that, when r is large, $\rho_i r$ is essentially the maximum number of type i customers the system can possibly have in steady state, because this would be the number of customers in the infinite-server system with no blocking. Thus, the assumption guarantees that it is feasible, at least in principle, to operate a system in a way such that, in the $r \to \infty$ limit, the steady-state blocking probability vanishes.

Consider the following function $L^{\square}(\bar{\boldsymbol{x}})$ defined on $\bar{\boldsymbol{x}}$ such that $\boldsymbol{x} \in \mathcal{X}^{\square}$ (and $x_{\boldsymbol{0}^s} \equiv h_s - \sum_{\boldsymbol{k}\in\mathcal{K}^s} h_{\boldsymbol{k}}$ for all s):

$$L^{\square}(\bar{\boldsymbol{x}}) = \sum_{\boldsymbol{k}\in\bar{\mathcal{K}}} x_{\boldsymbol{k}} \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/e], \tag{17}$$

where $c_{\boldsymbol{k}} \doteq \prod_i k_i!$, $0! = 1$. We then have

$$(\partial/\partial x_{\boldsymbol{k}}) L^{\square}(\bar{\boldsymbol{x}}) = \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}], \quad \boldsymbol{k} \in \bar{\mathcal{K}}. \tag{18}$$

For each $\boldsymbol{k} \in \bar{\mathcal{K}}$ the corresponding summand in the definition (17) of function $L^{\square}(\bar{\boldsymbol{x}})$ is strictly convex in $x_{\boldsymbol{k}}$; then, $L^{\square}(\bar{\boldsymbol{x}})$ is strictly convex on $\mathbb{R}_+^{|\bar{\mathcal{K}}|}$.

Consider the problem $\min_{\boldsymbol{x}\in\mathcal{X}^{\diamond}} L^{\square}(\bar{\boldsymbol{x}})$. It is the following convex optimization problem:

$$\min_{\bar{\boldsymbol{x}}\in\mathbb{R}_+^{|\bar{\mathcal{K}}|}} L^{\square}(\bar{\boldsymbol{x}}), \tag{19}$$

subject to

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, \quad \forall i, \tag{20}$$

$$\sum_{\boldsymbol{k}\in\bar{\mathcal{K}}^s} x_{\boldsymbol{k}} = h_s, \ s \in \mathcal{S}. \tag{21}$$

Denote by $\bar{\boldsymbol{x}}^{*,\Box}$ its unique optimal solution; of course, the corresponding $\boldsymbol{x}^{*,\Box} \in \mathcal{X}^{\diamond}$. Using (18) and Assumption 7 it is easy to see that $x_{\boldsymbol{k}}^{*,\Box} > 0$ for all $\boldsymbol{k} \in \bar{\mathcal{K}}$. There exist a vector of Lagrange multipliers $\boldsymbol{\nu}^{*,\Box} = (\nu_i^{*,\Box},\ i \in \mathcal{I})$ for the constraints (20) and Lagrange multipliers β_s^* for the constraints (21), such that $\bar{\boldsymbol{x}}^{*,\Box}$ solves problem

$$\min_{\bar{\boldsymbol{x}}\in\mathbb{R}_+^{|\bar{\mathcal{K}}|}} L^{\Box}(\bar{\boldsymbol{x}}) + \sum_i \nu_i^{*,\Box}(\rho_i - \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}) + \sum_s \beta_s^*(\sum_{\boldsymbol{k}\in\bar{\mathcal{K}}^s} x_{\boldsymbol{k}} - h_s).$$

We see that $\log[x_{\boldsymbol{k}}^{*,\Box} c_{\boldsymbol{k}}] - \sum_i \nu_i^{*,\Box} k_i + \beta_s^* = 0$, $\boldsymbol{k} \in \bar{\mathcal{K}}^s$. Therefore, $\bar{\boldsymbol{x}}^{*,\Box}$ has the product form

$$x_{\boldsymbol{k}}^{*,\Box} = \frac{1}{c_{\boldsymbol{k}}} \exp\left[-\beta_s^* + \sum_i k_i \nu_i^{*,\Box}\right] = \frac{e^{-\beta_s^*}}{c_{\boldsymbol{k}}} \exp\left[\sum_i k_i \nu_i^{*,\Box}\right], \quad \boldsymbol{k} \in \bar{\mathcal{K}}^s. \tag{22}$$

This in particular implies that Lagrange multipliers $\nu_i^{*,\Box}$, β_s^*, are unique. They can have any sign (not necessarily non-negative).

We obtain the following fact. *A point $\bar{\boldsymbol{x}}$, such that $\boldsymbol{x} \in \mathcal{X}^{\diamond}$, is the optimal solution to (19)-(21) (that is $\bar{\boldsymbol{x}} = \bar{\boldsymbol{x}}^{*,\Box}$) if and only if it has a product form representation (22) for some Lagrange multipliers $\nu_i^{*,\Box}$, β_s^*.* Furthermore, *$\boldsymbol{x}^{*,\Box}$ and $\boldsymbol{\nu}^{*,\Box}$ are equal to $\boldsymbol{x}^{*,\boldsymbol{a}}$ and $\boldsymbol{\nu}^{*,\boldsymbol{a}}$, respectively, defined for the infinite-server system in Section 2.2, with parameters $a_s = e^{-\beta_s^*}$.*

Our main result for the finite-server system is the following Proposition 8. (It is stated here informally. Formal statements are given in Lemmas 15 and 16.)

Proposition 8. *Suppose Assumption 7 holds. As $r \to \infty$, the limits of the fluid-scaled trajectories $\boldsymbol{x}^r(\cdot)$ will be referred to as fluid sample paths (FSP). Point $\boldsymbol{x} \in \mathcal{X}^{\Box}$ is an invariant point, if $\boldsymbol{x}(t) \equiv \boldsymbol{x}$ is an FSP. Then $\boldsymbol{x}^{*,\Box}$ is the unique invariant point $\boldsymbol{x}$, such that $x_{\boldsymbol{0}^s} > 0$ for all s (and therefore there is no blocking). Moreover, this invariant point is locally stable: $\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\Box}$, uniformly for all FSPs with $\boldsymbol{x}(0)$ sufficiently close to $\boldsymbol{x}^{*,\Box}$.*

In turn, Proposition 8 strongly suggests that the following asymptotic optimality property holds, which we present as

Conjecture 9. *Suppose Assumption 7 holds. Consider a sequence of systems under the GRAND-F algorithm, indexed by r, and let $\boldsymbol{x}^r(\infty)$ denote the random state of the fluid-scaled process in the stationary regime. Then, as $r \to \infty$, $\boldsymbol{x}^r(\infty) \implies \boldsymbol{x}^{*,\Box}$.*

If Conjecture 9 is correct, the GRAND-F algorithm is asymptotically optimal in the following sense. As long as Assumption 7 holds, i.e. the system has enough capacity to process all offered load (under ideal packing), then as $r \to \infty$, the steady-state blocking probability under GRAND-F vanishes. As discussed in Remark 6, GRAND-F can be viewed as an extension of PULL algorithm [16]. Therefore, Conjecture 9, if correct, can be viewed as an extension (to systems with packing constraints) of the asymptotic optimality of PULL.

3 Proof of Theorem 3

For any $\boldsymbol{k} \in \mathcal{K}^s$, as $a_s \downarrow 0$,

$$[-\log a_s]^{-1} x_{\boldsymbol{k}} \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/(e a_s)] - x_{\boldsymbol{k}} = [-\log a_s]^{-1} x_{\boldsymbol{k}}[\log x_{\boldsymbol{k}} + \log c_{\boldsymbol{k}} - 1] \to 0,$$

uniformly on any compact subset of non-negative $x_{\boldsymbol{k}}$. We have

$$L^{(\boldsymbol{a})}(\boldsymbol{x})/[-\log a_1] = \sum_s [-\log a_s]/[-\log a_1] \sum_{\boldsymbol{k}\in\mathcal{K}^s} [-\log a_s]^{-1} x_{\boldsymbol{k}} \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/(e a_s)].$$

Setting $a_s = \alpha^{\gamma_s}$ (which implies $[-\log a_s]/[-\log a_1] = \gamma_s/\gamma_1 = \gamma_s$), we see that, as $\alpha \downarrow 0$, $|L^{(\boldsymbol{a})}(\boldsymbol{x})/[-\log \alpha] - \sum_s \sum_{\boldsymbol{k}\in\mathcal{K}^s} \gamma_s x_{\boldsymbol{k}}| \to 0$, uniformly in $\boldsymbol{x} \in \mathcal{X}$. Therefore, $\boldsymbol{x}^{*,\boldsymbol{a}}$ must converge to $\mathcal{X}^*$.

Consider any sequence $\alpha \downarrow 0$. We will denote $b = -\log \alpha$. We will show that from any subsequence we can choose a further subsequence, along which we have convergence $\boldsymbol{x}^{*,\boldsymbol{a}} \to \boldsymbol{x}^*$, $\boldsymbol{\nu}^{*,\boldsymbol{a}}/b \to \boldsymbol{\eta}^*$, where $\boldsymbol{x}^* \in \mathcal{X}^*$ and $\boldsymbol{\eta}^* \in \mathcal{H}^*$.

Let a subsequence of α be fixed. Since $\boldsymbol{x}^{*,\boldsymbol{a}} \to \mathcal{X}^*$, we can and do choose a further subsequence along which $\boldsymbol{x}^{*,\boldsymbol{a}} \to \boldsymbol{x}^*$ for some fixed $\boldsymbol{x}^* \in \mathcal{X}^*$. Let us show that

$$\limsup_{\alpha\to 0} \sum_i k_i \nu_i^{*,\boldsymbol{a}}/b \le \gamma_s, \quad \forall \boldsymbol{k} \in \mathcal{K}^s, \tag{23}$$

$$\liminf_{\alpha\to 0} \nu_i^{*,\boldsymbol{a}}/b \ge 0, \quad \forall i. \tag{24}$$

From (15) we have:

$$x_{\boldsymbol{k}}^{*,\boldsymbol{a}} = \frac{1}{c_{\boldsymbol{k}}} \exp\left[b(\sum_i k_i \nu_i^{*,\boldsymbol{a}}/b - \gamma_s) \right], \quad \boldsymbol{k} \in \mathcal{K}^s. \tag{25}$$

If (23) would not hold for some $\boldsymbol{k} \in \mathcal{K}^s$, then by (25) we would have $\limsup x_{\boldsymbol{k}}^{*,\boldsymbol{a}} = \infty$ – a contradiction. Thus, (23) holds. Suppose now that (24) does not hold for some i, that is $\liminf \nu_i^{*,\boldsymbol{a}}/b < 0$. Pick an s and $\boldsymbol{k} \in \mathcal{K}^s$ such that $k_i \ge 1$ and $x_{\boldsymbol{k}}^* > 0$. Such s and $\boldsymbol{k}$ must exist, because $\sum_{\boldsymbol{k}} k_i x_{\boldsymbol{k}}^* = \rho_i$ (recall that $\boldsymbol{x}^* \in \mathcal{X}^*$). Since $x_{\boldsymbol{k}}^{*,\boldsymbol{a}} \to x_{\boldsymbol{k}}^* \in [0, \rho_i]$, we see from (25) that $\lim \sum_j k_j \nu_j^{*,\boldsymbol{a}}/b = \gamma^s$. Therefore,

$$\limsup \left[\sum_{j\ne i} k_j \nu_j^{*,\boldsymbol{a}}/b + (k_i - 1)\nu_i^{*,\boldsymbol{a}}/b \right] = \gamma^s - \liminf \nu_i^{*,\boldsymbol{a}}/b > \gamma^s;$$

but, this violates (23) for configuration $\boldsymbol{k} - \boldsymbol{e}_i$. Thus, (24) holds.

By (23)-(24), the sequence of $\boldsymbol{\nu}^{*,\boldsymbol{a}}/b$ is bounded. Then, we choose a further subsequence along which $\boldsymbol{\nu}^{*,\boldsymbol{a}}/b$ converges to some $\boldsymbol{\eta}^*$. For the pair $\boldsymbol{x}^*$ and $\boldsymbol{\eta}^*$, condition (6) is automatic, conditions (7)-(8) follow from (23)-(24), and condition (9) follows from (25). Therefore, $\boldsymbol{\eta}^* \in \mathcal{H}^*$. □

4 Fluid sample paths for the infinite-server system under GRAND(aZ). Proof of Theorem 2

In this section, we define fluid sample paths (FSP) for the system controlled by GRAND($\boldsymbol{a}Z$). FSPs arise as limits of the (fluid-scaled) trajectories $(1/r)\boldsymbol{X}^r(\cdot)$ as $r \to \infty$. Then we prove Theorem 2. The development in this section is a generalization to the heterogeneous system of the definitions and results given for the homogeneous system in Section 4 of [15]. The generalization is quite straightforward. However, we provide it here for completeness and, more importantly, as a preparation for the related argument used later in Section 5 for the finite-server system.

Let $\mathcal{M}$ denote the set of pairs $(\boldsymbol{k}, i)$ such that $\boldsymbol{k} \in \mathcal{K}$ and $\boldsymbol{k} - \boldsymbol{e}_i \in \bar{\mathcal{K}}$. Each pair $(\boldsymbol{k}, i)$ is associated with the "edge" $(\boldsymbol{k} - \boldsymbol{e}_i, \boldsymbol{k})$ connecting configurations $\boldsymbol{k} - \boldsymbol{e}_i$ and $\boldsymbol{k}$; often we refer to this edge as $(\boldsymbol{k}, i)$. By "arrival along the edge $(\boldsymbol{k}, i)$", we will mean placement of a type i customer into a server configuration $\boldsymbol{k} - \boldsymbol{e}_i$ to form configuration $\boldsymbol{k}$. Similarly, "departure along the edge $(\boldsymbol{k}, i)$" is a departure of a type-i customer from a server in configuration $\boldsymbol{k}$, which changes its configuration to $\boldsymbol{k} - \boldsymbol{e}_i$.

Without loss of generality, assume that the Markov process $X^r(\cdot)$ for each r is driven by the common set of primitive processes, defined as follows.

For each $(\boldsymbol{k}, i) \in \mathcal{M}$, consider an independent unit-rate Poisson process $\{\Pi_{\boldsymbol{k}i}(t),\ t \geq 0\}$, which drives departures along edge $(\boldsymbol{k}, i)$. Namely, let $D^r_{\boldsymbol{k}i}(t)$ denote the total number of departures along the edge $(\boldsymbol{k}, i)$ in $[0, t]$; then

$$D^r_{\boldsymbol{k}i}(t) = \Pi_{\boldsymbol{k}i}\left(\int_{\mathbf{0}}^t X^r_{\boldsymbol{k}}(s) k_i \mu_i ds\right). \tag{26}$$

The functional strong law of large numbers (FSLLN) holds:

$$\frac{1}{r}\Pi_{\boldsymbol{k}i}(rt) \to t, \quad u.o.c., \quad w.p.1. \tag{27}$$

For each $i \in \mathcal{I}$, consider an independent unit-rate Poisson process $\Pi_i(t),\ t \geq 0$, which drives exogenous arrivals of type i. Namely, let $A^r_i(t)$ denote the total number of type-i arrivals in $[0, t]$, then

$$A^r_i(t) = \Pi_i(\lambda_i rt). \tag{28}$$

Analogously to (27),

$$\frac{1}{r}\Pi_i(rt) \to t, \quad u.o.c., \quad w.p.1. \tag{29}$$

The random placement of new arrivals is constructed as follows. For each $i \in \mathcal{I}$, consider an i.i.d. sequence $\xi_i(1), \xi_i(2), \ldots$ of random variables, uniformly distributed in $[0, 1]$. Denote by $\mathcal{K}_i \doteq \{\boldsymbol{k} \in \bar{\mathcal{K}} \mid \boldsymbol{k} + \boldsymbol{e}_i \in \bar{\mathcal{K}}\}$ the subset of those configurations (including zero configurations) which can fit an additional type-i customer. The configurations $\boldsymbol{k} \in \mathcal{K}_i$ are indexed by $1, 2, \ldots, |\mathcal{K}_i|$ (in arbitrary fixed order). When the m-th (in time) customer of type i arrives in the system, it is assigned as follows. If $X^r_{(i)} = 0$, the customer is assigned to an empty server of an arbitrarily fixed type s, such that $\boldsymbol{e}_i \in \mathcal{K}^s$. Suppose $X^r_{(i)} \geq 1$. Then, the customer is assigned to a server in configuration $\boldsymbol{k}'$ indexed by 1 if

$$\xi_i(m) \in [0, X^r_{\boldsymbol{k}'}/X^r_{(i)}],$$

it is assigned to a server in configuration $\boldsymbol{k}''$ indexed by 2 if

$$\xi_i(m) \in (X^r_{\boldsymbol{k}''}/X^r_{(i)}, (X^r_{\boldsymbol{k}'} + X^r_{\boldsymbol{k}''})/X^r_{(i)}],$$

and so on. Denote

$$g^r_i(\sigma, \zeta) \doteq \sum_{m=1}^{\lfloor r\sigma \rfloor} I\{\xi_i(m) \leq \zeta\},$$

where $\sigma \geq 0$, $0 \leq \zeta \leq 1$. Obviously, from the strong law of large numbers and the monotonicity of $g^r_i(\sigma, \zeta)$ on both arguments, we have the FSLLN

$$g^r_i(\sigma, \zeta) \to \sigma\zeta, \quad \text{u.o.c.} \quad \text{w.p.1} \tag{30}$$

It is easy (and standard) to see that, for any r, w.p.1, the realization of the process $\{\boldsymbol{X}^r(t),\ t \geq 0\}$ is uniquely determined by the initial state $\boldsymbol{X}^r(0)$ and the realizations of the driving processes $\Pi_{\boldsymbol{k}i}(\cdot)$, $\Pi_i(\cdot)$ and $(\xi_i(1), \xi_i(2), \ldots)$.

If we denote by $A^r_{\boldsymbol{k}i}(t)$ the total number of arrivals allocated along edge $(\boldsymbol{k}, i)$ in $[0, t]$, we obviously have $\sum_{\boldsymbol{k} \in \mathcal{K}_i} A^r_{\boldsymbol{k}i}(t) = A^r_i(t),\ t \geq 0$, for each i.

In addition to

$$x^r_{\boldsymbol{k}}(t) = \frac{1}{r} X^r_{\boldsymbol{k}}(t),$$

we introduce other fluid-scaled quantities:

$$d^r_{\boldsymbol{k}i}(t) = \frac{1}{r} D^r_{\boldsymbol{k}i}(t), \quad a^r_{\boldsymbol{k}i}(t) = \frac{1}{r} A^r_{\boldsymbol{k}i}(t).$$

A set of locally Lipschitz continuous functions $[\{x_{\boldsymbol{k}}(\cdot),\ \boldsymbol{k}\in\mathcal{K}\},\{d_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\},\{a_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\}]$ on the time interval $[0,\infty)$ we call a *fluid sample path* (FSP), if there exist realizations of the primitive driving processes, satisfying conditions (27),(29) and (30) and a fixed subsequence of r, along which

$$
\begin{aligned}
&[\{x^r_{\boldsymbol{k}}(\cdot),\ \boldsymbol{k}\in\mathcal{K}\},\{d^r_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\},\{a^r_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\}]\rightarrow\\
&[\{x_{\boldsymbol{k}}(\cdot),\ \boldsymbol{k}\in\mathcal{K}\},\{d_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\},\{a_{\boldsymbol{k}i}(\cdot),\ (\boldsymbol{k},i)\in\mathcal{M}\}],\quad u.o.c.
\end{aligned}
\tag{31}
$$

For any FSP, all points $t>0$ are regular (see definition in Section 1.3), except a subset of zero Lebesgue measure.

Lemma 10. *Consider a sequence of fluid-scaled processes $\{\boldsymbol{x}^r(t),\ t\geq 0\}$ with fixed initial states $\boldsymbol{x}^r(0)$ such that $\boldsymbol{x}^r(0)\rightarrow\boldsymbol{x}(0)$. Then w.p.1, for any subsequence of r there exists a further subsequence of r, along which the convergence (31) holds, with the limit being an FSP.*

Proof is same as that of Lemma 5 in [15]. □

For an FSP, at a regular time point t, we denote $v_{\boldsymbol{k}i}(t)=(d/dt)a_{\boldsymbol{k}i}(t)$ and $w_{\boldsymbol{k}i}(t)=(d/dt)d_{\boldsymbol{k}i}(t)$. In other words, $v_{\boldsymbol{k}i}(t)$ and $w_{\boldsymbol{k}i}(t)$ are the rates of type-i "fluid" arrival and departure along edge $(\boldsymbol{k},i)$, respectively. Also denote: $y_i(t)=\sum_{\boldsymbol{k}}k_i x_{\boldsymbol{k}}(t)$, $z(t)=\sum_i y_i(t)$, $x_{\boldsymbol{0}^s}(t)=a_s z(t)$, and $x_{(i)}(t)=\sum_{\boldsymbol{k}\in\bar{\mathcal{K}}:\boldsymbol{k}+\boldsymbol{e}_i\in\bar{\mathcal{K}}}x_{\boldsymbol{k}}(t)$.

Lemma 11. *(i) An FSP satisfies the following properties at any regular point t:*

$$(d/dt)y_i(t)=\lambda_i-\mu_i y_i(t),\quad \forall i\in\mathcal{I},\tag{32}$$

$$w_{\boldsymbol{k}i}(t)=k_i\mu_i x_{\boldsymbol{k}}(t),\quad \forall(\boldsymbol{k},i)\in\mathcal{M},\tag{33}$$

$$x_{(i)}(t)>0\ \ \textit{implies}\ \ v_{\boldsymbol{k}i}(t)=\frac{x_{\boldsymbol{k}-\boldsymbol{e}_i}(t)}{x_{(i)}(t)}\lambda_i,\quad \forall(\boldsymbol{k},i)\in\mathcal{M},\tag{34}$$

$$\sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}}v_{\boldsymbol{k}i}(t)=\lambda_i,\quad \forall i\in\mathcal{I},\tag{35}$$

$$(d/dt)x_{\boldsymbol{k}}(t)=\left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}}v_{\boldsymbol{k}i}(t)-\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\bar{\mathcal{K}}}v_{\boldsymbol{k}+\boldsymbol{e}_i,i}(t)\right]-\left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}}w_{\boldsymbol{k}i}(t)-\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\bar{\mathcal{K}}}w_{\boldsymbol{k}+\boldsymbol{e}_i,i}(t)\right],\quad \forall\boldsymbol{k}\in\mathcal{K}.\tag{36}$$

Clearly, (32) implies

$$y_i(t)=\rho_i+(y_i(0)-\rho_i)e^{-\mu_i t},\quad t\geq 0,\quad \forall i\in\mathcal{I}.\tag{37}$$

(ii) Moreover, an FSP with $\boldsymbol{x}(0)\in\mathcal{X}$ satisfies the following stronger conditions:

$$y_i(t)\equiv\rho_i,\quad \forall i\in\mathcal{I},\tag{38}$$

$$z(t)\equiv 1,\quad x_{\boldsymbol{0}^s}(t)\equiv a_s,\quad x_{(i)}(t)\geq\sum_{s:\ \boldsymbol{e}_i\in\mathcal{K}^s}a_s,\ \forall i\in\mathcal{I};\tag{39}$$

at any regular point t,

$$v_{\boldsymbol{k}i}(t)=\frac{x_{\boldsymbol{k}-\boldsymbol{e}_i}(t)}{x_{(i)}(t)}\lambda_i,\quad \forall(\boldsymbol{k},i)\in\mathcal{M},\tag{40}$$

$$\sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}}w_{\boldsymbol{k}i}(t)=\lambda_i,\quad \forall i\in\mathcal{I}.\tag{41}$$

Proof. (i) Given the convergence (31), which defines an FSP, all the stated properties except (34) are nothing but the limit versions of the flow conservations laws. Property (34) follows from the construction of the random assignment, the continuity of $\boldsymbol{x}(t)$, and (30). We omit further details.
(ii) If $\boldsymbol{x}(0)\in\mathcal{X}$, which implies $y_i(0)=\rho_i$ for each i, property (38) (and then (39) as well) follows from (37). Then, (34) strengthens to (40), and (41) is verified directly using (33). □

Lemma 12. *For any FSP with* $\boldsymbol{x}(0) \in \mathcal{X}$,

$$\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\boldsymbol{a}}, \tag{42}$$

and the convergence is uniform across all such FSPs.

Proof. Given that $x_{\mathbf{0}^s}(t) \equiv a_s$ and $\sum_{\boldsymbol{k}} x_{\boldsymbol{k}}(t) \le 1$, we have $x_{(i)}(t) \le 1 + \sum_s a_s$, hence $v_{\boldsymbol{k}i}(t) \ge x_{\boldsymbol{k}}(t)\lambda_i/(1 + \sum_s a_s)$. From here, we obtain the following fact: for any $\boldsymbol{k}$ and any $\delta > 0$ there exists $\delta_1 > 0$ such that for all $t \ge \delta$, $x_{\boldsymbol{k}}(t) \ge \delta_1$. The proof is by contradiction. Consider a $\boldsymbol{k}$, say $\boldsymbol{k} \in \bar{\mathcal{K}}^s$, that is a minimal counterexample; necessarily, $\boldsymbol{k} \ne \mathbf{0}^s$. Pick any $\delta > 0$ and then the corresponding $\delta_1 > 0$ such that the statement holds for any $\boldsymbol{k}' \in \bar{\mathcal{K}}^s$, $\boldsymbol{k}' < \boldsymbol{k}$. (Here $\boldsymbol{k}' < \boldsymbol{k}$ means that $\boldsymbol{k}'_i \le \boldsymbol{k}_i$, $\forall i$, and $\boldsymbol{k}' \ne \boldsymbol{k}$.) We observe from (36) that for any regular $t \ge \delta$, $(d/dt)x_{\boldsymbol{k}}(t) > \delta_2 > 0$ as long as $x_{\boldsymbol{k}}(t) \le \delta_3$, for some positive constants δ_2, δ_3. Since this holds for an arbitrarily small $\delta > 0$ (with $\delta_1, \delta_2, \delta_3$ depending on it), we see that the statement is true for $\boldsymbol{k}$.

In particular, we see that $x_{\boldsymbol{k}}(t) > 0$ for all $t > 0$ and all $\boldsymbol{k}$. Note also that all $t > 0$ are regular points (because all $w_{\boldsymbol{k}i}$ and $v_{\boldsymbol{k}i}$ are bounded continuous in $\boldsymbol{x}$).

To prove the lemma, it will suffice to show that:
(a) if $\boldsymbol{x}(t) \ne \boldsymbol{x}^{*,\boldsymbol{a}}$ and $x_{\boldsymbol{k}}(t) > 0$ for all $\boldsymbol{k} \in \mathcal{K}$, then $(d/dt)L^{(\boldsymbol{a})}(\boldsymbol{x}(t)) < 0$; and, moreover,
(b) the derivative is bounded away from zero as long as $\|\boldsymbol{x}(t) - \boldsymbol{x}^{*,\boldsymbol{a}}\|$ is bounded away from zero.
Let us denote by $\Xi(\boldsymbol{x})$ the derivative $(d/dt)L^{(\boldsymbol{a})}(\boldsymbol{x}(t))$ at a given point $\boldsymbol{x}(t) = \boldsymbol{x}$; in the rest of the proof we study the function $\Xi(\boldsymbol{x})$ on $\mathcal{X}$, and therefore drop the time index t. Suppose all components $x_{\boldsymbol{k}} > 0$. From (33), (35), (40), and (41), we have:

$$w_{\boldsymbol{k}i} = k_i\mu_i x_{\boldsymbol{k}} = k_i\mu_i x_{\boldsymbol{k}} \sum_{\boldsymbol{k}':(\boldsymbol{k}',i)\in\mathcal{M}} \frac{x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{x_{(i)}}, \tag{43}$$

$$v_{\boldsymbol{k}'i} = \frac{x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{x_{(i)}}\lambda_i = \frac{x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{x_{(i)}} \sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}} k_i\mu_i x_{\boldsymbol{k}}. \tag{44}$$

Expressions (43) and (44) can be interpreted as follows. For any ordered pair of edges $(\boldsymbol{k}, i)$ and $(\boldsymbol{k}', i)$, we can assume that the part $k_i\mu_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}/x_{(i)}$ of the total departure rate $k_i\mu_i x_{\boldsymbol{k}}$ along $(\boldsymbol{k}, i)$ is "allocated back" as a part of the arrival rate along $(\boldsymbol{k}', i)$. Using (11), the contribution of these "coupled" departure/arrival rates for the ordered pair of edges $(\boldsymbol{k}, i)$ and $(\boldsymbol{k}', i)$ into the derivative $\Xi(\boldsymbol{x})$ is

$$\xi_{\boldsymbol{k},\boldsymbol{k}',i} = [\log(k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}) - \log(k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i})]\,\frac{k_i\mu_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{x_{(i)}}.$$

This expression is valid even when either $\boldsymbol{k} - \boldsymbol{e}_i = \mathbf{0}^s$ or $\boldsymbol{k}' - \boldsymbol{e}_i = \mathbf{0}^s$ for some s. This is because $x_{\mathbf{0}^s}(t) = a_s$ when $\boldsymbol{x} \in \mathcal{X}$, and therefore by convention (12), formula (11) is valid for all $\boldsymbol{k} \in \bar{\mathcal{K}}$. We have:

$$\xi_{\boldsymbol{k},\boldsymbol{k}',i} + \xi_{\boldsymbol{k}',\boldsymbol{k},i} = (\mu_i/x_{(i)})[\log(k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}) - \log(k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i})][k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i} - k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}] \le 0,$$

and the inequality is strict unless $k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'} = k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}$. We obtain

$$\Xi(\boldsymbol{x}) = \sum_i \sum_{\boldsymbol{k},\boldsymbol{k}'} [\xi_{\boldsymbol{k},\boldsymbol{k}',i} + \xi_{\boldsymbol{k}',\boldsymbol{k},i}]. \tag{45}$$

Therefore, $\Xi(\boldsymbol{x}) < 0$ unless $\boldsymbol{x}$ has a product form representation (15), which in turn is equivalent to $\boldsymbol{x} = \boldsymbol{x}^{*,a}$.

So far the function $\Xi(\boldsymbol{x})$ in (45) was defined for $\boldsymbol{x} \in \mathcal{X}$ with all $x_{\boldsymbol{k}} > 0$. Let us adopt a convention that $\Xi(\boldsymbol{x}) = -\infty$ for $\boldsymbol{x} \in \mathcal{X}$ with at least one $x_{\boldsymbol{k}} = 0$. Then, it is easy to verify that $\Xi(\boldsymbol{x})$ is continuous on the entire set $\mathcal{X}$.

It remains to show that for any $\delta_2 > 0$ there exists $\delta_3 > 0$ such that conditions $\boldsymbol{x} \in \mathcal{X}$ and $L^{(\boldsymbol{a})}(\boldsymbol{x}) - L^{(\boldsymbol{a})}(\boldsymbol{x}^{*,\boldsymbol{a}}) \ge \delta_2$ imply $\Xi(\boldsymbol{x}) \le -\delta_3$. This is indeed true, because otherwise there would exist $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x} \ne \boldsymbol{x}^{*,\boldsymbol{a}}$, such that $\Xi(\boldsymbol{x}) = 0$, which is, again, equivalent to $\boldsymbol{x} = \boldsymbol{x}^{*,\boldsymbol{a}}$. □

From Lemma 12 we easily obtain Theorem 2; see the proof of Theorem 3 in Section 4 of [15].

As in [15], we also have the following generalization of Lemma 12, showing FSP uniform convergence for arbitrary initial states, not necessarily $\boldsymbol{x}(0) \in \mathcal{X}$.

Lemma 13. *For any compact $A \in \mathbb{R}_+^{|\mathcal{K}|}$, the convergence*

$$\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\boldsymbol{a}} \tag{46}$$

holds uniformly in all FSPs with $\boldsymbol{x}(0) \in A$.

Proof repeats that of Lemma 8 in [15] almost verbatim. The only adjustments are:
1) Starting any fixed time $\tau > 0$, we have $0 < a_1 \le x_{\boldsymbol{0}^s}(t),\ \forall s$, and $x_{(i)}(t) \le a_2 < \infty,\ \forall i$, for some constants a_1, a_2, uniformly on all FSPs with $\boldsymbol{x}(0) \in A$;
2) $L^{(\boldsymbol{a})}$ replaces $L^{(a)}$;
3) $f(\boldsymbol{k}) = (\partial/\partial x_{\boldsymbol{k}})L^{(\boldsymbol{a})}(x) = \log[x_{\boldsymbol{k}} c_{\boldsymbol{k}}/a_s],\ \boldsymbol{k} \in \mathcal{K}^s$. □

5 GRAND-F: Local stability of FSPs

The construction of the Markov process $X^r(\cdot)$ under GRAND-F is the same as in Section 4 for GRAND($\boldsymbol{a}Z$), except now, when $X^r_{(i)} = 0$, an arriving type i customer is blocked. Consequently, we no longer have the identity $\sum_{\boldsymbol{k}\in\mathcal{K}_i} A^r_{\boldsymbol{k}i}(t) = A^r_i(t),\ t \ge 0$, for each i. Instead,

$$A^r_i(t) - \sum_{\boldsymbol{k}\in\mathcal{K}_i} A^r_{\boldsymbol{k}i}(t),\ t \ge 0,$$

is non-negative non-decreasing function, giving the number of blocked type i customers by time t.

The definition of an FSP and Lemma 10 hold as is. All points $t > 0$ are regular, except for a subset of zero Lebesgue measure. The analog of Lemma 11 is the following

Lemma 14. *(i) An FSP satisfies the following properties at any regular point t:*

$$\sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}} v_{\boldsymbol{k}i}(t) \le \lambda_i, \quad \forall i \in \mathcal{I}, \tag{47}$$

$$(d/dt)y_i(t) = \sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}} v_{\boldsymbol{k}i}(t) - \mu_i y_i(t), \quad \forall i \in \mathcal{I}, \tag{48}$$

$$w_{\boldsymbol{k}i}(t) = k_i \mu_i x_{\boldsymbol{k}}(t), \quad \forall(\boldsymbol{k}, i) \in \mathcal{M}, \tag{49}$$

$$x_{(i)}(t) > 0 \quad \textit{implies} \quad \sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}} v_{\boldsymbol{k}i}(t) = \lambda_i,\ \forall i \in \mathcal{I}, \quad \textit{and} \quad v_{\boldsymbol{k}i}(t) = \frac{x_{\boldsymbol{k}-\boldsymbol{e}_i}(t)}{x_{(i)}(t)}\lambda_i,\ \forall(\boldsymbol{k}, i) \in \mathcal{M}, \tag{50}$$

$$(d/dt)x_{\boldsymbol{k}}(t) = \left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} v_{\boldsymbol{k}i}(t) - \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\bar{\mathcal{K}}} v_{\boldsymbol{k}+\boldsymbol{e}_i,i}(t)\right] - \left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} w_{\boldsymbol{k}i}(t) - \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\bar{\mathcal{K}}} w_{\boldsymbol{k}+\boldsymbol{e}_i,i}(t)\right], \quad \forall \boldsymbol{k} \in \bar{\mathcal{K}}. \tag{51}$$

(ii) Moreover, an FSP with $\boldsymbol{x}(0) \in \mathcal{X}^\diamond$, $x_{(i)}(0) > 0,\ \forall i$, satisfies the following stronger conditions for all sufficiently small $t > 0$:

$$y_i(t) \equiv \rho_i, \quad \forall i \in \mathcal{I}, \tag{52}$$

$$z(t) \equiv 1, \quad x_{\boldsymbol{0}^s}(t) \equiv a_s,\ \forall s, \quad x_{(i)}(t) \ge \min_s a_s,\ \forall i \in \mathcal{I}; \tag{53}$$

if t is regular,

$$v_{\boldsymbol{k}i}(t) = \frac{x_{\boldsymbol{k}-\boldsymbol{e}_i}(t)}{x_{(i)}(t)}\lambda_i, \quad \forall(\boldsymbol{k}, i) \in \mathcal{M}, \tag{54}$$

$$\sum_{\boldsymbol{k}:(\boldsymbol{k},i)\in\mathcal{M}} w_{\boldsymbol{k}i}(t) = \lambda_i, \quad \forall i \in \mathcal{I}. \tag{55}$$

Proof. (i) Given the convergence (31) defining an FSP, all the stated properties except (50), are nothing but the limit versions of the flow conservations laws. Property (50) follows from the construction of the random assignment, the continuity of $\boldsymbol{x}(t)$, and (30). We omit further details.
(ii) If $\boldsymbol{x}(0) \in \mathcal{X}^{\diamond}$, which implies $y_i(0) = \rho_i$ for each i, property (52) (and then (53) as well) follows from (48) and (50). Then, (55) is verified directly using (49). Finally, (54) follows from (50). $\Box$

Lemma 15. *There exists $\epsilon > 0$, such that, uniformly on FSPs with initial states $\boldsymbol{x}(0) \in \mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon\}$,*

$$\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\Box}, \quad t \to \infty. \tag{56}$$

FSP $\boldsymbol{x}(t) \equiv \boldsymbol{x}^{*,\Box}$ is the unique invariant FSP, satisfying conditions $x_{\boldsymbol{0}^s}(0) > 0,\ \forall s$.

Proof. We can assume (without loss of generality) that ϵ is small enough so that $x_{\boldsymbol{k}}(0) > 0,\ \forall \boldsymbol{k} \in \bar{\mathcal{K}}$. In particular, at $t = 0$, the condition $x_{\boldsymbol{0}^s}(t) > 0,\ \forall s$, holds. Obviously, until the first time $\tau > 0$ when this condition is violated ($\tau = \infty$ if it is never violated), we have $y_i(t) = \rho_i,\ \forall i$. It is also easy to see that all time ponts $0 < t < \tau$ are regular and such that $x_{\boldsymbol{k}}(t) > 0,\ \forall \boldsymbol{k} \in \bar{\mathcal{K}}$. Denote by $\Xi(\bar{\boldsymbol{x}})$ the derivative $(d/dt)L^{\Box}(\bar{\boldsymbol{x}}(t))$ at a given point $\boldsymbol{x}(t) = \boldsymbol{x}$. Then, expressions (43) and (44) for $w_{\boldsymbol{k}i}$ and $v_{\boldsymbol{k}'i}$ hold for our system, and can be interpreted the same way. (Recall, however, that now the components $x_{\boldsymbol{0}^s}$ are *not* constant, and therefore their derivatives do depend on the rates $w_{\boldsymbol{0}^s+\boldsymbol{e}_i,i}$ and $v_{\boldsymbol{0}^s+\boldsymbol{e}_i,i}$.) Then the expression for $\Xi(\bar{\boldsymbol{x}})$ has exactly same form as expression (45) for $\Xi(\boldsymbol{x})$ in Section 4:

$$\Xi(\bar{\boldsymbol{x}}) = \sum_i \sum_{(\boldsymbol{k},i),(\boldsymbol{k}',i)} = (\mu_i/x_{(i)})[\log(k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}) - \log(k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i})][k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i} - k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}] \le 0. \tag{57}$$

The inequality in (57) is strict unless $k'_i x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'} = k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}$ for all pairs of edges $(\boldsymbol{k}, i)$ and $(\boldsymbol{k}', i)$. Therefore, $\Xi(\bar{\boldsymbol{x}}) < 0$ unless $\bar{\boldsymbol{x}}$ has a product form representation (22), which in turn is equivalent to $\boldsymbol{x} = \boldsymbol{x}^{*,\Box}$.

Function $\Xi(\bar{\boldsymbol{x}})$ is continuous in a neighborhood of $\boldsymbol{x}^{*,\Box}$ (and in fact at any point such that $x_{\boldsymbol{k}} > 0,\ \forall \boldsymbol{k} \in \bar{\mathcal{K}}$). Choose $\epsilon_1 > 0$ small enough so that $x_{\boldsymbol{k}} > 0,\ \boldsymbol{k} \in \bar{\mathcal{K}}$, for all $\boldsymbol{x} \in \mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon_1\}$. Then choose $\delta > 0$ such that condition $L^{\Box}(\bar{\boldsymbol{x}}) - L^{\Box}(\bar{\boldsymbol{x}}^{*,\Box}) \le \delta$ (along with $\boldsymbol{x} \in \mathcal{X}^{\diamond}$) implies $\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| < \epsilon_1$. Finally, choose $\epsilon > 0$ small enough so that the maximum of $L^{\Box}(\bar{\boldsymbol{x}}) - L^{\Box}(\bar{\boldsymbol{x}}^{*,\Box})$ over the set $\mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon\}$ is less than δ. We see that a trajectory with $\boldsymbol{x}(0) \in \mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon\}$ cannot escape from the set $\mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon_1\}$, and therefore $x_{\boldsymbol{k}}(t) > 0,\ \boldsymbol{k} \in \bar{\mathcal{K}}$, for all $t \ge 0$. Then, the convergence (56) holds, and it is uniform on $\boldsymbol{x}(0) \in \mathcal{X}^{\diamond} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon\}$, because, for any $0 < \delta_1 < \delta$, $\Xi(\bar{\boldsymbol{x}})$ is negative and bounded away from zero for all $\boldsymbol{x} \in \mathcal{X}^{\diamond} \cap \{\delta_1 \le L^{\Box}(\bar{\boldsymbol{x}}) - L^{\Box}(\bar{\boldsymbol{x}}^{*,\Box}) \le \delta\}$.

It is a corollary from the above argument, that there cannot be an invariant FSP $\boldsymbol{x}(t) \equiv \boldsymbol{x}(0)$ with $x_{\boldsymbol{0}^s}(0) > 0,\ \forall s$, unless $\boldsymbol{x}(0) = \boldsymbol{x}^{*,\Box}$. (Indeed, $\boldsymbol{x}(0) \in \mathcal{X}^{\diamond}$ necessarily, because if $y_i(0) \ne \rho_i$ then $y_i(t)$ cannot be constant. Then $\boldsymbol{x}(0) = \boldsymbol{x}^{*,\Box}$, because otherwise $L^{\Box}(\bar{\boldsymbol{x}}(t))$ cannot be constant.) This proves the second statement of the lemma. $\Box$

Lemma 16. *There exists $\epsilon > 0$, such that, uniformly on FSPs with initial states $\boldsymbol{x}(0) \in \mathcal{X}^{\Box} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\Box}\| \le \epsilon\}$,*

$$\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\Box}, \quad t \to \infty. \tag{58}$$

Proof is a slightly generalized version of that of Lemma 15. That proof considers FSPs that stay within $\mathcal{X}^{\diamond}$, uses the continuity of $\Xi(\bar{\boldsymbol{x}})$, and the fact that for $\boldsymbol{x} \in \mathcal{X}^{\diamond}$ in a small neighborhood of $\boldsymbol{x}^{*,\Box}$, $\Xi(\bar{\boldsymbol{x}}) < 0$ unless $\boldsymbol{x} = \boldsymbol{x}^{*,\Box}$. But, $\Xi(\bar{\boldsymbol{x}})$ is continuous in a neighborhood of $\boldsymbol{x}^{*,\Box}$ (or any point such that $x_{\boldsymbol{k}} > 0,\ \forall \boldsymbol{k} \in \bar{\mathcal{K}}$), not necessarily restricted to $\mathcal{X}^{\diamond}$. In addition, we know that as long as $x_{\boldsymbol{0}^s}(t) > 0,\ \forall s$, each $y_i(t)$ satisfies ODE $(d/dt)[y_i(t) - \rho_i] = -\mu_i[y_i(t) - \rho_i]$, and therefore

$$y_i(t) - \rho_i = (y_i(0) - \rho_i)e^{-\mu_i t}. \tag{59}$$

Using these observations, the adjustment of the proof of Lemma 15 is as follows. We choose small $\epsilon_1 > 0$, then $\delta > 0$, then $\epsilon > 0$, exactly as in that proof. Then, using the continuity of $\Xi(\bar{\boldsymbol{x}})$, along with (59), we can choose a sufficiently small $\epsilon_2 > 0$, so that a trajectory with $\boldsymbol{x}(0) \in \{|y_i - \rho_i| \le \epsilon_2, \ \forall i\} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\square}\| \le \epsilon\}$ cannot escape from the set $\{|y_i - \rho_i| \le \epsilon_2, \ \forall i\} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\square}\| \le \epsilon_1\}$. Then, the convergence (58) holds, and it is uniform on $\boldsymbol{x}(0) \in \{|y_i - \rho_i| \le \epsilon_2, \ \forall i\} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\square}\| \le \epsilon\}$, because, for any $0 < \delta_1 < \delta$, there exists a small $\epsilon_2' > 0$, such that $\Xi(\bar{\boldsymbol{x}})$ is negative and bounded away from zero for all $\boldsymbol{x} \in \{|y_i - \rho_i| \le \epsilon_2', \ \forall i\} \cap \{\delta_1 \le L^{\square}(\bar{\boldsymbol{x}}) - L^{\square}(\bar{\boldsymbol{x}}^{*,\square}) \le \delta\}$. (Note that the time for FSPs starting in $\{|y_i - \rho_i| \le \epsilon_2, \ \forall i\} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\square}\| \le \epsilon\}$ to reach set $\{|y_i - \rho_i| \le \epsilon_2', \ \forall i\} \cap \{\|\boldsymbol{x} - \boldsymbol{x}^{*,\square}\| \le \epsilon\}$ is uniformly bounded due to (59).) $\square$

5.1 Comments on Conjecture 9, local stability, and fixed point argument

Lemmas 15 and 16 formally state properties described informally in Proposition 8. The sequence of steady-states $\boldsymbol{x}^r(\infty)$ is obviously tight. It is easy to see that its any subsequential limit in distribution, $\boldsymbol{x}(\infty)$, is such that $y_i(\infty) \le \rho_i, \ \forall i$, w.p.1. This is because, by comparison with the infinite-server system, $Y_i^r(\infty)$ is stochastically dominated by a Poisson random variable with mean $\rho_i r$. Furthermore, again by comparison with the infinite-server system, any FSP with

$$\boldsymbol{x}(0) \in \mathcal{X}^{\square,\le} \equiv \{\boldsymbol{x} \in \mathcal{X}^{\square} \mid \sum_s \sum_{\boldsymbol{k} \in \mathcal{K}^s} k_i x_{\boldsymbol{k}} \le \rho_i, \ \forall i \in \mathcal{I}\}$$

stays in $\mathcal{X}^{\square,\le}$ at all times t. Given these facts, if we would have the (analogous to Lemma 12) uniform convergence property

$$\boldsymbol{x}(t) \to \boldsymbol{x}^{*,\square}, \quad \forall \boldsymbol{x}(0) \in \mathcal{X}^{\square,\le}, \tag{60}$$

this would prove Conjecture 9 (by the same argument as in the proof of Theorem 2). Unfortunately, the uniform convergence (60) does *not* hold for a general finite-server system. It is very easy to construct a counterexample (e.g., for a system with one server type with the configuration set shown of Fig. 1(b) in [15]) such that there exists an invariant FSP $\boldsymbol{x}(t) \equiv \boldsymbol{x}^*$, "sitting" at a suboptimal point $\boldsymbol{x}^* \ne \boldsymbol{x}^{*,\square}$, such that $y_i^* < \rho_i, \ \forall i$, and therefore such that there is non-zero fraction of customers of each type being blocked. (In fact, we believe that a stronger property holds for such a counterexample: the sequence of processes $\boldsymbol{x}^r(\cdot)$ converges *in distribution* to the invariant FSP $\boldsymbol{x}(t) \equiv \boldsymbol{x}^*$.) This, of course, does not imply that Conjecture 9 is wrong – it just shows that there is no hope of proving Conjecture 9 based on fluid scale considerations alone.

Lemmas 15 and 16 show FSP local stability at the optimal point $\boldsymbol{x}^{*,\square}$, and the fact that $\boldsymbol{x}^{*,\square}$ is the only invariant point at which there is no blocking. This strongly suggests that Conjecture 9 is correct, even though, as discussed above, it is insufficient for its proof. Still, we note that the local stability is a substantially stronger property than a typical "fixed point" argument which is used to "guess" asymptotic properties like our Conjecture 9. In our case a "fixed point" argument would go as follows: as $r \to \infty$, assume that steady-state distributions of server states are asymptotically independent; further assume that a subsequential limit of the marginal distribution of a server state is such that the server is empty with non-zero probability; under these assumptions, find the set of (limiting) marginal distributions (for each server type), which would remain invariant ("fixed") over time; in our case, this argument leads to finding that the only such possible set of marginal distributions is such that the system must be "sitting" at the point $\boldsymbol{x}^{*,\square}$, equal to the one defined in this paper. Note that, in essence, the above argument is nothing else but the statement that $\boldsymbol{x}^{*,\square}$ is the unique invariant point (at which there is no blocking) for FSPs, while local stability properties in Lemmas 15 and 16 are much stronger.

6 Discussion

Proving Conjecture 9 for the finite-server system under GRAND-F is a very interesting and challenging subject of future work. As discussed in Section 5.1, fluid-scale analysis alone cannot be sufficient for such a proof, because there may exists sub-optimal points, which are invariant for the FSPs.

The local stability results for the finite-server system with blocking (Proposition 8, Lemmas 15 and 16) hold for other variants of the finite-server system as well. Indeed, these results and their proofs only concern with the system behavior in the vicinity of equilibrium point, where there are always available servers for any customer type. Suppose now that we have a system in which customers are queued instead of blocking when there are no available servers for them (or a system where both blocking and queueing are possible). Then the local stability results still apply for this system, *as long as the assignment rule coincides with GRAND-F when there are servers available to arrivals.* Further, this suggests that Conjecture 9 is also valid for such other variants of the finite-server system, under appropiate versions of GRAND-F. In fact, recall that GRAND-F, as defined in this paper, itself can be viewed as an extension of PULL algorithm [16] to systems with packing constraints. PULL algorithm has been defined and proved to be asymptotically optimal for very general systems with queueing and/or blocking (but without packing constraints).

The results of this paper further highlight the universality of GRAND algorithm. For example, Best Fit type algorithms are applicable only to the special case of vector packing constraints, where the underlying notion of a customer "fitting best into the remaining space" at a server makes sense. When packing constraints are more general, Best Fit is not applicable, while GRAND is. Furthermore, inherently, Best Fit requires precise information about the current state of each server – this can be a substantial disadvantage in practical large-scale systems. GRAND, on the other hand, only needs to know whether a given customer fits into a given server or not; this allows a very efficient practical implementation (as discussed in detail in Remark 6). It is possible that versions of Best Fit may perform better than GRAND for systems with vector packing constraints. Paper [5] provides some evidence of that. (Although, the algorithm studied in [5] is not a "pure" Best Fit, but a Best Fit *with randomization*, a mixture, in a sense, of Best Fit and GRAND.) Studying versions of Best Fit is an interesting subject; it is outside the scope of this paper, which is focused on general packing constraints. First Fit is another approach to packing; algorithms of this type use fixed preordering of servers and place each customer into the first one where it can fit. Such algorithms are easily implementable and apply to general packing constraints. Note that GRAND can be viewed as a First Fit with random uniform reordering of servers before each customer placement. If the order of servers has to chosen and fixed a priori, as "pure" First Fit requires, the question arises on how to do it when the servers are heterogeneous, as in our model. Exploring variants of First Fit may be another subject of future research.

References

[1] N. Bansal, A. Caprara, M. Sviridenko. A New Approximation Method for Set Covering Problems, with Applications to Multidimensional Bin Packing. *SIAM J. Comput.*, 2009, Vol.39, No.4, pp. 1256-1278.

[2] M. Bramson, Y. Lu and B. Prabhakar. Asymptotic independence of queues under randomized load balancing. *Queueing Systems*, 2012, Vol.71, pp. 247-292.

[3] M. Bramson, Y. Lu and B. Prabhakar. Decay of tails at equlibrium for fifo join the shortest queue networks. *The Annals of Applied Probability*, 2013, Vol.23, pp. 1841-1878.

[4] J. Csirik, D. S. Johnson, C. Kenyon, J. B. Orlin, P. W. Shor, and R. R. Weber. On the Sum-of-Squares Algorithm for Bin Packing. *J.ACM*, 2006, Vol.53, pp.1-65.

[5] G. Ghaderi, Y. Zhong and R. Srikant. Asymptotic optimality of BestFit for stochastic bin packing. *SIGMETRICS-2014*, pp.64-66. DOI 10.1145/2667522.2667543

[6] A. Gulati, A. Holler, M. Ji, G. Shanmuganathan, C. Waldspurger, X. Zhu. VMware Distributed Resource Management: Design, Implementation and Lessons Learned. *VMware Technical Journal*, 2012, Vol.1, No.1, pp. 45-64. http://labs.vmware.com/publications/vmware-technical-journal

[7] Y. Guo, A. L. Stolyar, A. Walid. Shadow-routing based dynamic algorithms for Virtual Machine placement in a network cloud. *INFOCOM-2013.* http://ect.bell-labs.com/who/stolyar/publications/gpd-vm-paper-inf.pdf

[8] V. Gupta, A. Radovanovic. Online Stochastic Bin Packing. Preprint, 2012. http://arxiv.org/abs/1211.2687

[9] Y. Lu, Q. Xe, G. Kilot, A. Geller, J. Larus and A. Greenberg. Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, 2011, Vol. 89, pp. 1057-1071.

[10] S. T. Maguluri, R. Srikant, L.Ying. Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters. *INFOCOM-2012*.

[11] S. T. Maguluri, R. Srikant. Scheduling Jobs with Unknown Duration in Clouds. *INFOCOM-2013*.

[12] M. Mitzenmacher. The power of two choices in randomized load balancing. *IEEE Transactions on Parallel and Distributed Systems*, 2001, Vol. 12, No. 10, pp. 1094-1104.

[13] A. L. Stolyar. An infinite server system with general packing constraints. *Operations Research*, 2013, Vol.61, No.5, pp. 1200-1217.

[14] A. L. Stolyar, Y. Zhong. A large-scale service system with packing constraints: Minimizing the number of occupied servers. *SIGMETRICS-2013*. http://arxiv.org/abs/1212.0875

[15] A. L. Stolyar, Y. Zhong. Asymptotic optimality of a greedy randomized algorithm in a large-scale service system with general packing constraints. *Queueing Systems*, 2015, Vol.79, No.2, pp. 117-143. DOI 10.1007/s11134-014-9414-x

[16] A. L. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems*, 2015, Vol.80, No.4, pp.341-361. DOI 10.1007/s11134-015-9448-8

[17] A. L. Stolyar, Y. Zhong. A service system with packing constraints: Greedy randomized algorithm achieving sublinear in scale optimality gap. Preprint, Nov. 2015. Submitted. http://arxiv.org/abs/1511.03241

[18] N. Vvedenskaya, R. Dobrushin and F. Karpelevich. (1996). Queueing system with selection of the shortest of two queues: an asymptotic approach. *Problems of Information Transmission*, 1996, Vol. 32, pp. 20-34.

[19] Q. Xie, X. Dong, Y. Lu and R. Srikant. Power of d Choices for Large-Scale Bin Packing: A Loss Model. *SIGMETRICS-2015*, pp.321-334.